

Under consideration for publication in Knowledge and Information
Systems

Discovering Topic Structures of a Temporally Evolving Document Corpus

Adham Beykikhoshk[†], Ognjen Arandjelović[‡], Dinh Phung[†], Svetha Venkatesh[†]

[†] Pattern Recognition and Data Analytics Centre, Deakin University, Australia

{abeyki, dinh.phung, svetha.venkatesh}@deakin.edu.au

[‡] School of Computer Science, University of St Andrews, UK

ognjen.arandjelovic@gmail.com

Abstract. In this paper we describe a novel framework for the discovery of the topical content of a data corpus, and the tracking of its complex structural changes across the temporal dimension. In contrast to previous work our model does not impose a prior on the rate at which documents are added to the corpus nor does it adopt the Markovian assumption which overly restricts the type of changes that the model can capture. Our key technical contribution is a framework based on (i) discretization of time into epochs, (ii) epoch-wise topic discovery using a hierarchical Dirichlet process-

Received xxx

Revised xxx

Accepted xxx

based model, and (iii) a temporal similarity graph which allows for the modelling of complex topic changes: emergence and disappearance, evolution, splitting, and merging. The power of the proposed framework is demonstrated on two medical literature corpora concerned with the autism spectrum disorder (ASD) and the metabolic syndrome (MetS) – both increasingly important research subjects with significant social and healthcare consequences. In addition to the collected ASD and metabolic syndrome literature corpora which we made freely available, our contribution also includes an extensive empirical analysis of the proposed framework. We describe a detailed and careful examination of the effects that our algorithms’s free parameters have on its output, and discuss the significance of the findings both in the context of the practical application of our algorithm as well as in the context of the existing body of work on temporal topic analysis. Our quantitative analysis is followed by several qualitative case studies highly relevant to the current research on ASD and MetS, on which our algorithm is shown to capture well the actual developments in these fields.

Keywords: Data mining; non-parametric; Bayesian, autism, metabolic syndrome.

1. Introduction

In the last decade and a half so-called topic modelling has emerged as a powerful statistical paradigm for the automatic semantic analysis of large collections of documents. Topic models as their name suggests can be seen as formalizations of the colloquial understanding of ‘topics’ addressed in a piece of text. More specifically, in this context a topic becomes a probability distribution over a fixed vocabulary of words (or more generally terms). Thus an example of a particular topic may be:

$$\begin{array}{ccccccc}
0.2 & 0.15 & 0.11 & 0.06 & \dots & & \\
\textit{autistic} & \textit{child} & \textit{education} & \textit{needs} & \dots & & \\
\hline
& \text{vocabulary size} & & & & &
\end{array} \tag{1}$$

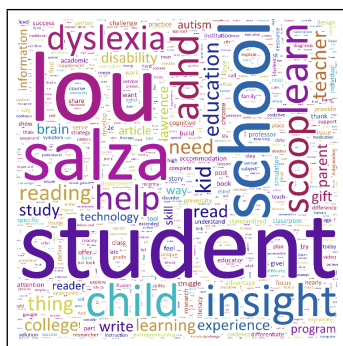
where the upper row contains probabilities which correspond to the vocabulary words shown in the bottom row. Using higher order semantic understanding a human interpreting this formal representation of a topic may describe it as capturing a discussion of educational needs of children with ASD although it should be noted that such interpretation may not always be straightforward [16].

Most research on topic modelling to date has focused on the analysis in the context of *static* corpora, that is, document collections which do not possess a temporal structure. In such collections documents are said to be interchangeable. The key techniques dominating this domain are Bayesian non-parametric inference algorithms and the latent Dirichlet allocation (LDA) in particular, first described by Blei *et al.* [15] and subsequently extended in a variety of ways. Indeed at the bottom-most level the present paper uses a model based on the hierarchical Dirichlet process (HDP) which is one of the aforementioned extensions. Both LDA and HDP are explained in some detail in Section 3.1.

However, in many problems of practical significance, it is not only the instantaneous topic structure that is of interest – the change in this structure over time too often conveys important information and insight. For this reason in recent years the problem of temporal topic modelling has been attracting an increasing amount of research attention [8, 9]. Indeed the focus of the present paper is on temporally changing corpora. At the heart of the method that we describe is an automatically constructed and temporally constrained graph su-



(b) Example topic 2



(d) Example topic 4

Fig. 1. An illustration of typical topics discovered by the proposed method in a single epoch, shown as size-coded word-clouds – a larger font indicates a proportionally more probable term (different colours are merely used for easier visualization and encode no information pertaining to the corresponding terms themselves). Topic 1 can be readily related to the persistent myth of a link between childhood vaccination and autism development as well as related issues concerning the effects of thimerosal and mercury [33], Topic 2 to the incident involving Chris Tuttle, a Wegmans employee with Asperger’s syndrome who attracted media attention in November 2013, Topic 3 to medical literature on autism, children, and brain development, and Topic 4 to the schooling of children with learning disabilities.

perimposed on topics extracted from short time epochs within which the corpus can be considered as being static.

In particular the work described herein extends our contribution first described in [8] (also see [10, 11] for related prior work). Retaining the same structural framework, in the present paper we analyse in detail the effects of different pruning parameters in the construction of the graph superimposed over the extracted topics. Specifically we consider different choices of inter-topic distance measures and the process of selecting an appropriate pruning threshold given a specific distance measure. In addition, we describe extended experiments on the collection of abstracts of scholarly articles on the autism spectrum disorder (ASD), gathered by ourselves and made publicly available, as well as additional experiments on a newly gathered collection of abstracts of scholarly articles in the highly active sphere of work on the metabolic syndrome. This new collection of documents will also be made publicly available.

The remainder of the present paper is structured as follows. In the next section we review relevant previous work, both on topic modelling in general as well as on temporal modelling which is the focus of our work. Then in Section 3 we first describe the key prerequisite modelling techniques which our contributions builds upon (Section 3.1), followed by the detail on the proposed modelling framework (Section 3.2). Section 3.2 introduces our main technical contributions. Specifically these are our general temporal model based on what we term the temporal similarity graph, and the two key aspects involved in the construction of this graph: the choice of an inter-topic similarity measure (Section 3.2.1) and a graph pruning process (Section 3.2.2). The proposed method is extensively evaluated in Section 4. Section 4.1 motivates our focus on biomedical documents used in the evaluation with Section 4.1.1 summarizing previous work in the spe-

cific domain of text mining, Section 4.1.2 provides a brief synopsis of the medical conditions the two specific document data sets we used for evaluation relate to, and Section 4.1.3 describes the data sets themselves, the manner in which data was collected, as well as the pre-processing thereof. Section 4.2 details different inter-topic similarity measures we compared. Our experiments and the associated findings and discussion are presented in Section 4.3, with Section 4.3.1 focusing on quantitative elements of the evaluation and Section 4.3.2 on qualitative analysis. Finally, the contributions and the key findings of the present work are summarized in Section 5.

2. Previous work

Owing to its growing popularity in recent years, the existing literature on probabilistic topic modelling is already vast. Hence a comprehensive survey of the field is out of the scope of the present paper. Herein we overview the broad and most influential research directions, with a particular focus on techniques of direct relevance to the methodology described in the proposed paper. In particular we direct our attention first to latent topic models which have dominated the field in the last decade, and then on biomedical text mining, given the application domain within which our framework is evaluated in Section 4.

Topic models in modern machine learning are often described as latent probabilistic models. The attribute ‘latent’ is intended to capture the nature of inferred topics – these are hidden variables in the sense that they are not explicit in the observable data itself. Arguably the sought after “true” topic structure is also neither objective nor accessible even in principle. On the other hand, the attribute ‘probabilistic’ conveys the inherent aspect of the aforementioned mod-

els whereby modelling imperfections as well as ambiguities in data are handled through the use of probability distributions which readily accommodate uncertainty. Therefore we start our discussion with the simplest latent topic models which formalize the key ideas in the field and underlie much of the subsequent, more complex models.

2.1. Latent topic models

An important early topic modelling approach comes in the form of so-called latent semantic indexing (LSI) [19] which remains popular. Two notable limitations of LSI are its inability to deal effectively with polysemy and to produce an explicit description of the latent space. A probabilistic improvement overcomes these by explicitly characterizing the latent space with semantic topics, and by employing a probabilistic generative model that addresses the polysemy problem [31]. Nevertheless, probabilistic LSI is prone to parameter overfitting caused by an uncontrolled growth in the number of parameters as the document corpus is increased. In addition, the necessary assignment of probabilities to documents is a nontrivial task [15].

The recently proposed latent Dirichlet allocation (LDA) method [15] overcomes the overfitting problem by adopting a Bayesian framework and a generative process at the document level. While LDA has quickly become a standard tool for topic modelling, it too experiences challenges when applied on real-world data. In particular, being a parametric model the number of desired output topics has to be specified in advance. The HDP model as the nonparametric counterpart of LDA was introduced by Teh *et al.* [55] and addressed this limitation by using a hierarchical Dirichlet process (as opposed to a Dirichlet distribution) as

the prior on topics. Therefore, each document is modelled using an infinite mixture model, allowing the data to inform the complexity of the model and infer the number of resulting topics automatically. We discuss this model in further detail in Section 3.

2.1.1. Temporal topic modelling

A notable limitation of most models described in the existing literature lies in their assumption that the data corpus is static; this includes those based on LDA mentioned previously, or the hierarchical Dirichlet process described in detail in the next section. Here the term ‘static’ is used to describe the lack of any associated temporal (or indeed sequential) information associated with the documents in a corpus – the documents are said to be exchangeable [14]. However, in many practical applications documents are added to the corpus in a temporal manner and their ordering has significance i.e. the documents are non-exchangeable. As a consequence, the topical structure of the corpus changes over time. For example, the corpus of scholarly literature on a particular subject is a growing corpus which by its very nature exhibits significant changes in its topic structure over time [22, 21]: new ideas emerge, old ideas are refined, novel discoveries result in multiple ideas being related to one another thereby forming more complex concepts or a single idea multifurcating into different ‘sub-ideas’ which are thereafter investigated with some degree of independence etc. A good example from a different realm can be readily found in the corpus of social media contributions, such as Twitter [9]. With an even faster pace here too complex topic changes can be observed, with novel topics of conversation being instigated by e.g. ‘real-world’ events (such as epidemics, terrorist attacks, or developments

in the world of popular culture), changed by the contributions of other users, split into new topics, merged with others etc.

The assumption made by all previous work, and indeed adopted by us, is that documents are not exchangeable at large temporal scales but are at short time scales, thus treating the corpus as *temporally locally static*. The scale at which this assumption can be considered as valid is clearly application and corpus dependent, and is an important consideration. Indeed the present paper investigates this in detail.

The existing work on temporal topic modelling can be divided into two groups of approaches both of which can be based on parametric [14, 58, 59] or nonparametric [44, 62] techniques, the former suffering from the limitation that they contain free parameters which must be set *a priori*. Methods of the first group discretize time into epochs, apply a static topic model to each epoch, and by making the Markovian assumption relate the parameters of each epoch’s topic model to those of the epochs adjacent to it in time [14, 58, 44, 62]. While the approach we propose in this paper adopts the idea of time discretization, it diverges in its other features from this group of methods thereafter. In particular, instead of employing the Markovian assumption we describe a novel structure in form of a temporal similarity graph, which gives our method greater flexibility, as described in detail in the next section. The second group of methods in the literature regard document time-stamps as observations of a continuous random variable [59, 20]. This assumption severely limits the type of topic changes which can be described. For example, as opposed to our model, these models are not capable of describing the evolution of topics, or their splitting and merging, and are rather constrained to tracking simple topic popularity.

3. Proposed framework

In this section we present the technical contribution of the present work. We begin by reviewing the relevant theory underlying Bayesian mixture modelling, LDA, and HDP that plays a central role in the proposed framework. Then we turn our attention to the novel contribution of our work and explain how the aforementioned Bayesian techniques are extended to deal with temporally varying document corpora.

3.1. Bayesian mixture models

In recent years mixture models have become popular choices for the modelling of so-called heterogeneous data. In this context heterogeneity is taken to mean that observable data is generated by more than one process (source). One of the key challenges in the analysis of heterogeneous data lies in the lack of observability of the correspondence between specific data points and their sources i.e. it is not known which data source generated which data. Usually it is also the case that the number of sources is not known either [46]. Mixture models, and in particular mixture models enveloped within a Bayesian framework, have distinct advantages over alternative approaches as we shall explain shortly.

3.1.1. Finite mixture modelling

Finite mixture models rely on the assumption that the observed data is generated by K clusters, each cluster being associated with the parameter ϕ_k and underlain by the probability density function $f(\cdot|\phi_k)$. An observation x is assumed to be generated by first choosing a cluster k with probability π_k followed by a random draw from the corresponding distribution described by ϕ_k . Therefore the process

can be summarized by the following:

$$p(x|\pi_{1:K}, \phi_{1:K}) = \sum_{k=1}^K \pi_k f(x|\phi_k). \quad (2)$$

In a Bayesian setting the model parameters (i.e. mixing proportions $\pi_{1:K}$ and component parameters $\phi_{1:k}$) are further endowed by priors. Typically the symmetric Dirichlet distribution is placed on top of $\pi_{1:K}$ and a prior on $\phi_{1:K}$ conjugate with $f(\cdot|\phi_k)$ chosen for computational convenience.

3.1.2. Latent Dirichlet allocation

In the previous section we described how to model a group of data points with a Bayesian finite mixture model. Latent Dirichlet allocation adds a level of hierarchy on the mixing proportions to allow for the modelling of data points in groups that share a set of components.

Following the consensus in the literature we adopt the terminology used in the analysis of textual data (which is the context in which LDA was originally proposed [15]) and hereafter interexchangably refer to data points as words, their groups as documents, and mixture components as topics. The technical term ‘topic’ can be interpreted as formalizing and abstracting the colloquial notion of a topic which is understood at a higher semantic level. Therefore the modelling framework of LDA can be described by the following generative process:

$$\phi_{1:K} \sim H, \quad (3)$$

$$\pi_j \sim \text{Dirichlet}(\alpha), \quad (4)$$

$$z_{ji}|\pi_j \sim \pi_j, \quad (5)$$

$$x_{ji}|z_{ji}, \phi_{1:K} \sim F(\phi_{z_{ji}}), \quad (6)$$

where H is the base distribution of topics, α the hyperparameter of the prior

on the distribution of topics within a document, π_j the distribution of topics in document j , and z_{ji} the topic corresponding to the i -th word in the j -th document. The corresponding model likelihood is:

$$p(w_{ji}|\alpha) = \int_{\pi_j} \int_{\phi_{1:K}} \sum_{k=1}^K \pi_{jk} f(x|\phi_k) dP(\pi_j) dP(\phi_{1:K}), \quad (7)$$

Approximation techniques such as MCMC [27] and Variation Bayes [15] methods can be used for posterior inference.

3.1.3. Infinite mixture modelling

As mentioned earlier, LDA requires the number of topics to be fixed in advanced which is a serious limitation in practice. Choosing the number of topics is usually performed by examining how well the model fits a set of held-out documents. However, if a previously unseen topic has contributed in generating the held-out data, LDA is not able to infer correct parameters of that topic.

Bayesian non-parametric (BNP) methods place priors on the infinite-dimensional space of probability distributions and provide an elegant solution to this problem. Dirichlet Process (DP) [23] as the non-parametric counterpart of the Dirichlet distribution and the building block of BNP allows for the model to accommodate a potentially infinite number of mixture components. The generative likelihood for a data point x in infinite mixture model is:

$$p(x|\pi_{1:\infty}, \phi_{1:\infty}) = \sum_{k=1}^{\infty} \pi_k f(x|\phi_k). \quad (8)$$

DP (γ, H) is defined as a distribution of a random probability measure G over a measurable space (Θ, \mathcal{B}) , such that for any finite measurable partition (A_1, A_2, \dots, A_r) of Θ the random vector $(G(A_1), \dots, G(A_r))$ is a Dirichlet distribution with parameters $(\gamma H(A_1), \dots, \gamma H(A_r))$. A DP generates imperfect atomic copies of its base measure H with a variance governed by its concen-

tration parameter γ . An alternative view of the DP emerges from the so-called stick-breaking process which adopts a constructive approach using a sequence of discrete draws [50]. Specifically, if $G \sim \text{DP}(\gamma, H)$ then $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ where $\phi_k \stackrel{iid}{\sim} H$ and $\beta = (\beta_k)_{k=1}^{\infty}$ is the vector of weights obtained by a stick-breaking process that is $\beta_k = v_k \prod_{l=1}^{k-1} (1 - v_l)$ and $v_l \stackrel{iid}{\sim} \text{Beta}(1, \gamma)$.

Owing to the discrete nature and infinite dimensionality of its draws, the DP is a highly useful prior for Bayesian mixture models. By associating different mixture components with atoms ϕ_k of the stick-breaking process, and assuming $x_i | \phi_k \stackrel{iid}{\sim} f(x_i | \phi_k)$ where $f(\cdot)$ is the likelihood kernel of the mixing components, we can formulate the infinite Bayesian mixture model or Dirichlet process mixture model (DPM). Approximate methods are used for posterior inference [43].

3.1.4. Hierarchical Dirichlet process mixture models

While DPM is suitable for non-parametric clustering of exchangeable data in a single group, many real-world problems are more appropriately modelled as comprising multiple groups of exchangeable data. In such cases it is usually desirable to model the observations of different groups jointly, allowing them to share their generative clusters. This idea is known as “sharing the statistical strength” and it is naturally obtained by hierarchical architecture in Bayesian modelling.

Consider a collection of documents. DPM models each group with an infinite number of topics. However, it is desired for multiple group-level DPMs to share their clusters. Amongst different ways of linking group-level DPMs, HDP [55] offers an interesting solution whereby base measures of group-level DPs are drawn from a corpus-level DP. In this way the atoms of the corpus-level DP (i.e. topics in our case) are shared across the documents. Formally, if $\mathbf{x} = \{\mathbf{x}_1, \dots, \mathbf{x}_J\}$ is a

document collection where $\mathbf{x}_j = \{x_{j1}, \dots, x_{jN_j}\}$ is the j -th document comprising N_j words:

$$G_0 | \gamma, H \sim \text{DP}(\gamma, H) \quad (9)$$

$$G_j | \alpha_0, G_0 \sim \text{DP}(\alpha_0, G_0) \quad (10)$$

$$\theta_{ji} | G_j \sim G_j \quad (11)$$

$$x_{ji} | \theta_{ji} \sim F(\cdot | \theta_{ji}) \quad (12)$$

This is illustrated schematically in Figure 2(a). Since G_j is drawn from a DP with base measure G_0 , it takes the same support as G_0 . Also the parameters of the group-level mixture components, θ_{ji} , share their values with the corpus-level DP support on $\{\phi_1, \phi_2, \dots\}$. Therefore G_j can be equivalently expressed using the stick-breaking process as $G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$ where $\pi_j | \alpha_0, \gamma \sim \text{DP}(\alpha_0, \gamma)$ [54]. The posterior for θ_{ji} has been shown to follow a Chinese restaurant franchise process which can be used to develop inference algorithms based on Gibbs sampling [55].

3.2. Modelling topic evolution over time

Hitherto, the discussion in this section focused on the modelling of static document corpora. We now show how our work builds on top of these ideas in the existing literature and in particular how the described HDP-based model can be applied to the analysis of temporal topic changes in a *longitudinal* data corpus.

Like some of the previous work in this area we begin by discretizing time and dividing the literature corpus by time into *epochs*. Each epoch spans a certain contiguous time period and has associated with it all documents with timestamps within this period. The duration of epochs should be sufficiently short to allow

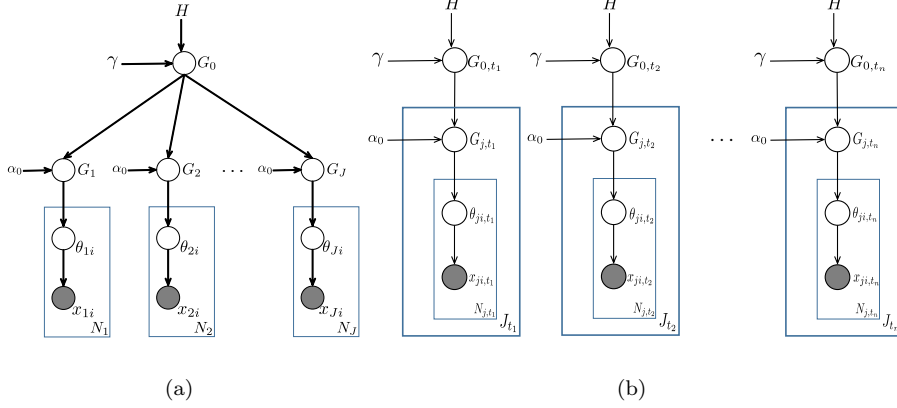


Fig. 2. (a) Graphical model representation of HDP. Each box represents one document whose observed data (words) is shown shaded. Unshaded nodes represent latent variables. An observed datum x_{ji} is assigned to a latent mixture component parameterized by θ_{ji} . γ and α are the concentration parameters and H is the corpus-level base measure. (b) Graphical model representation of the proposed framework. The corpus is temporally divided into t_n epochs and each epoch modelled using an HDP (outer boxes). Different epochs' HDPs share their corpus-level DP and hyperparameters.

the corresponding document subset to be treated as a static collection; we shall discuss this issue in more detail in the next section.

Each epoch is then modelled separately using an HDP, with models corresponding to different epochs sharing their hyperparameters and the corpus-level base measure. Hence if n is the number of epochs, we obtain n sets of topics $\phi = \{\phi_{t_1}, \dots, \phi_{t_n}\}$ where $\phi_t = \{\theta_{1,t}, \dots, \theta_{K_t,t}\}$ is the set of topics that describe epoch t , and K_t their number (which is inferred automatically, as described previously). This is illustrated in Figure 2(b). In the next section we describe how given an inter-topic similarity measure the evolution of different topics across epochs can be tracked.

3.2.1. *Measuring topic similarity*

Our goal now is to track changes in the topical structure of a data corpus over time. The simplest changes of interest include the emergence of new topics, and the disappearance of others. More subtly, we are also interested in how a specific topic changes, that is, how it evolves over time in terms of the contributions of different words it comprises. Lastly, our aim is to be able to extract and model complex structural changes of the underlying topic content which result from the interaction of older topics. Specifically, topics, which can be thought of as collections of memes, can merge to form new topics or indeed split into more nuanced memetic collections. This information can provide valuable insight into the refinement of ideas and findings in the scientific community, effected by new research and accumulating evidence.

The key idea behind our approach stems from the observation that while topics may change significantly over time, providing that the duration of epochs is chosen appropriately, the change between successive epochs is limited. Therefore we infer the continuity of a topic in one epoch by relating it to all topics in the immediately subsequent epoch which are sufficiently similar to it under a suitable similarity measure. This can be seen to lead naturally to a similarity graph representation whose nodes correspond to topics and whose edges link those topics in two epochs which are related. Formally, the weight of the directed edge that links $\phi_{j,t}$, the j -th topic in epoch t , and $\phi_{k,t+1}$ is set equal to $\rho(\phi_{j,t}, \phi_{k,t+1})$ where ρ is a similarity measure. Given that in our HDP-based model each topic is represented by a probability distribution, similarity measures such the Bhattacharyya coefficient [12], the Jensen-Shannon or Kullback-Leibler divergences [38, 3], and the Hellinger [30] or resistor-average distances [47], for

example, are all readily adapted for use in the proposed framework as we shall demonstrate in the next section.

A conceptual illustration of a small section of a similarity graph is shown in Figure 3. It shows three consecutive time epochs $t - 1$, t , and $t + 1$ and a selection of topics in these epochs. Graph edge weight (i.e. inter-topic similarity) is encoded by the thickness of the line representing the edge – the thicker the line, the more similar the corresponding topics are. In constructing a similarity graph we use a threshold to eliminate automatically weak edges, retaining only the connections between sufficiently similar topics in adjacent epochs. This readily allows us to detect the disappearance of a particular topic, the emergence of new topics, as well as the splitting or merging of different topics as follows:

– **New topic emergence:**

If a node does not have any edges incident to it, the corresponding topic is taken as having emerged in the associated epoch; for example, in Figure 3 the topic ϕ_{j+2} can be seen to emerge during the epoch with the timestamp t .

– **Topic disappearance:**

If no edges originate from a node, the corresponding topic is taken to vanish in the associated epoch; for example, in Figure 3 the topic ϕ_j can be seen to disappear during the epoch with the timestamp t .

– **Topic evolution:**

When exactly one edge originates from a node in one epoch and it is the only edge incident to a node in the following epoch, the topic is understood as having evolved in the sense that its memetic content (captured by the probability distribution over the underlying vocabulary) may have changed;

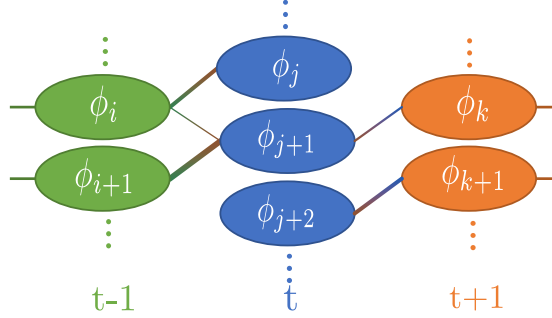


Fig. 3. Conceptual illustration of the proposed similarity graph that models topic dynamics over time. A node corresponds to a topic in a specific epoch; edge weights are equal to the corresponding topic similarities.

for example, in Figure 3 the topic ϕ_{j+2} evolves into the topic ϕ_{k+1} between the epochs with the timestamps t and $t + 1$.

– **Topic splitting:**

If more than a single edge originates from a node, we interpret this as the corresponding topic splitting into multiple topics between the corresponding epochs and the successive epoch; for example, in Figure 3 the topic ϕ_i splits into topics ϕ_j and ϕ_{j+1} between the epochs with the timestamps $t - 1$ and t .

– **Topic merging:**

If more than a single edge is incident to a node, the topics of the nodes from which the edges originate are understood to have interacted by merging to form a new topic; for example, in Figure 3 between the epochs with the timestamps $t - 1$ and t the topics ϕ_i and ϕ_{i+1} merge to form ϕ_{j+1} .

Lastly, observe that just as our understanding of topics at a higher semantic level would allow, the proposed framework readily models complex structural changes which involve a topic concurrently undergoing merging and splitting.

For example, the topic labelled ϕ_{j+1} in Figure 3 is created through the merging of topic ϕ_{i+1} and a split offshoot of the topic ϕ_i .

3.2.2. Automatic temporal similarity graph construction

In our previous work [8, 9] the temporal similarity graph is built in two stages. In the first stage the graph is connected fully in the sense that all pairs of topics in successive epochs are connected by edges. Then, the graph is pruned using a similarity threshold t_s . In other words any edge corresponding to an inter-topic similarity lesser than t_s is removed from the graph.

A major limitation of the described pruning step is that the similarity threshold t_s is not readily interpretable and the framework provides little insight as to how the threshold should be chosen. In addition, considering that the threshold value inherently depends on the similarity measure which is used, it is not clear how two inter-topic similarity measures may be compared i.e. how to control for the threshold in the presence of a changing similarity metric which underlies it. Hence in the present paper we describe an alternative strategy which employs a more meaningful and more interpretable manner of pruning. Firstly we consider all inter-topic similarities present in the initial fully connected graph and extract the empirical estimate of the corresponding cumulative density function (CDF). Examples of CDFs obtained using three similarity metrics on a typical epoch in our data (please see the next section for empirical analysis) are shown for illustration in Figure 4. The very different functional forms of the three CDFs shown reflect our previous observation that pruning on the basis of a similarity threshold is inherently dependent on the employed similarity measure, which complicates any comparative analysis. Instead of using a similarity threshold herein we prune the graph based on the operating point on the relevant CDF. In other words if

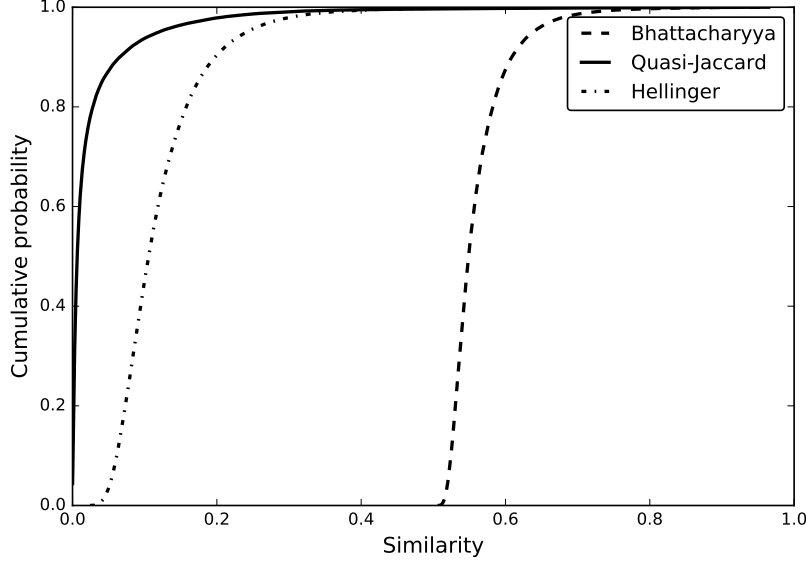


Fig. 4. Empirical estimates of the cumulative density function (CDF) of inter-topic similarities corresponding to fully connected temporal similarity graph constructed over the same set of topics but using three different similarity measures (please see Section 4.2). Such CDFs are used in the proposed construction of the final temporal similarity graph, that is, in the pruning process used to derive it from the precedent, fully connected graph.

F_ρ is the CDF corresponding to a specific initial, fully connected graph formed using a particular similarity measure, and $\zeta \in [0, 1]$ the CDF operating point, we prune the edge between topics $\phi_{j,t}$ and $\phi_{k,t+1}$ iff $\rho(\phi_{j,t}, \phi_{k,t+1}) < F_\rho^{-1}(\zeta)$.

4. Experimental evaluation

Having introduced the main technical contribution of our work we now analyse the performance of the proposed framework empirically on two large real-world data sets.

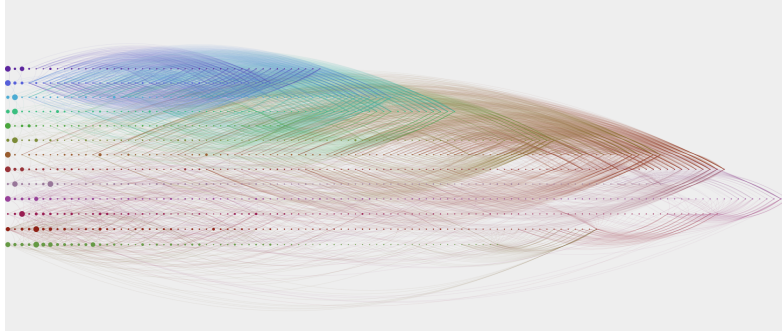
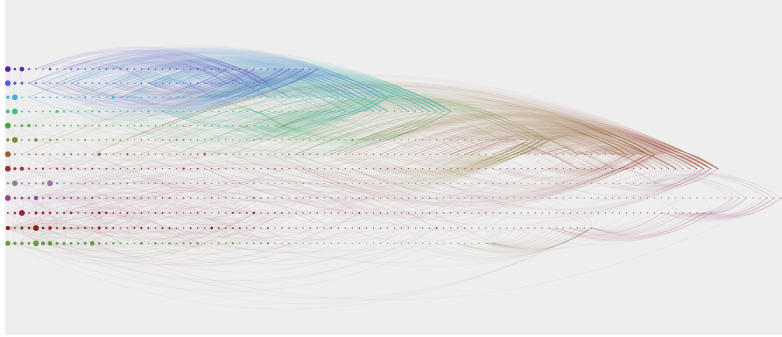
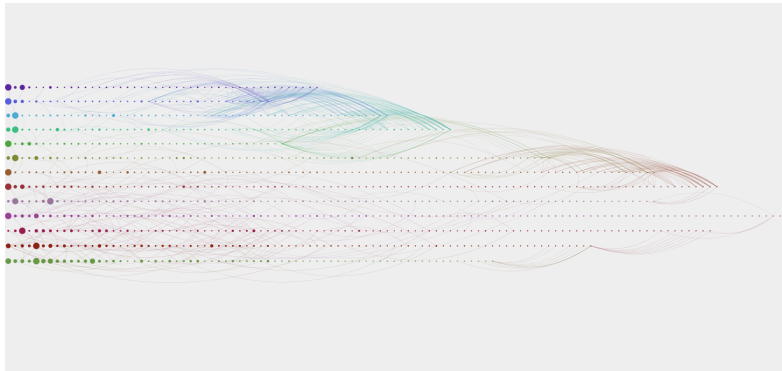
(a) $\zeta = 0.9$ (b) $\zeta = 0.95$ (c) $\zeta = 0.99$

Fig. 5. Examples of temporal similarity graphs. All three graphs correspond to the same data corpus and are formed over the same set of topics using a similarity measure based on the Hellinger distance (please see Section 4.2) but using different CDF operating points $\zeta = 0.9, 0.95, 0.99$. The horizontal axis denotes time with each column of topics (dots) corresponding to a specific epoch. The size of dots representing topics codes for their popularity in the corresponding epoch. As expected, increasing ζ results in an increase in sparsity of the final

4.1. Evaluation data

In Section 2 we highlighted the corpus of published peer-reviewed scientific literature as a particularly attractive data source for the application of temporal topic models. This corpus exhibits rapid growth at an accelerating pace, and it is by its very nature characterized by high fluency of ideas. These ideas can be represented well by topics in the sense in which this term is used in the present article. Hence for the evaluation of the method proposed in the preceding section we selected two highly active research areas in the realm of biomedicine.

For the sake of completeness we next briefly survey the most influential work on the analysis of biomedical texts, then outline the key reasons for our choice of the specific research areas we focus on, and then proceed with a description of the adopted experimental methodology, a presentation of the most significant results emerging from our analysis, and the associated discussion.

4.1.1. *Biomedical text mining*

Most previous work on text-based knowledge discovery in biomedicine to date has focused on (i) the tagging of names of entities such as genes, proteins, and diseases [51], (ii) the discovery of relationships between different entities e.g. functional associations between genes [45], or (iii) the extraction of information pertaining to events such as gene expression or protein binding [52].

The idea that the medical literature could be mined for new knowledge is typically attributed to Swanson [53]. For example by manually examining medical literature databases he hypothesised that dietary fish oil could be beneficial for Raynaud’s syndrome patients, which was later confirmed by experimental evidence. Work that followed sought to develop statistical methods which would

make this process automatic. Most approaches adopted the use of term frequencies and co-occurrences using dictionaries such as Medical Subject Headings (MeSH) [48].

Most existing work on biomedical knowledge discovery is based on what may be described as traditional data mining techniques (neural networks, support vector machines etc); comprehensive surveys can be found in [35, 52]. The application of state-of-the-art Bayesian methods in this domain is scarce. Amongst the notable exceptions is the work by Blei *et al.* who showed how latent Dirichlet allocation (LDA) can be used to facilitate the process of hypothesis generation in the context of genetics [13]. Arnold *et al.* used a similar approach to demonstrate that abstract topic space representation is effective in patient-specific case retrieval [4]. In their later work they introduced a temporal model which learns topic trends and showed that the inferred topics and their temporal patterns correlate with valid clinical events and their sequences [5]. Wu *et al.* used LDA for gene-drug relationship ranking [61].

4.1.2. Autism spectrum disorder and the metabolic syndrome

In this paper we evaluate our topic discovery framework on two corpora of scholarly papers. The first corpus is of abstracts of papers concerning the autism spectrum disorder, and the other of abstracts of papers related to research on the metabolic syndrome. These specific research areas are chosen for several reasons. Firstly, they concern medical issues of major practical importance – they affect (directly or indirectly) a large number of people and impose a significant financial cost both to the society as a whole and to those affected. Secondly, the understanding of mechanisms underlying both conditions have proven to pose a significant intellectual challenge. Consequently, the dominant ideas regarding

the underlying causative mechanisms, modes of treatment and their efficacy etc. are continually changing, experiencing both refinement as well as more abrupt paradigm shifts. These aspects make the chosen areas of research highly suitable for the evaluation of the framework described in the present work.

The autism spectrum disorder Autism spectrum disorder is a life-long neurodevelopmental disorder with poorly understood causes on the one hand, and a wide range of potential treatments supported by little evidence on the other. The disorder is characterized by severe impairments in social interaction, communication, and in some cases cognitive abilities, and typically begins in infancy or at the very latest by the age of three. ASD is recognized as comprising an aetiologically and clinically heterogeneous group of conditions whose diagnosis remains to be based solely on the complex behavioural phenotype [41]. According to the definition in the latest version (5th edition) of the Diagnostic and Statistical Manual of Mental Disorders, the autism spectrum disorder includes disorders which were previously diagnosed with more specificity as autism, Asperger syndrome, Rett syndrome, childhood disintegrative disorder, and ‘pervasive developmental disorder not otherwise specified’ [1]. Current evidence suggests that approximately 0.5-0.6% of the population is afflicted by ASD though the actual diagnosis rate is on the increase due to the broadening diagnostic criteria [6]. The condition is usually detected in early childhood when an abnormal lack of social reciprocity is observed.

Although the last few decades have seen significant progress in the study of ASD, the still relatively poorly understood aetiology of the condition, its phenotypical heterogeneity [37], and stigma associated with mental conditions [26], have all contributed to the penetration of beliefs, and behavioural and educa-

tional interventions which are often questionable [60] and poorly supported by evidence (e.g. gluten-free and casein-free diets, and cognitive behavioural therapy [18]), and sometimes outright in conflict with science [56]. For example a recent review of early intensive behavioural and developmental interventions for young children with ASD found 1 existing study as being of good quality, 10 as fair quality, and 23 as poor quality [60]. From the public policy point of view, understanding the practices and beliefs of parents and carers of ASD-affected individuals is crucial, yet often lacking [29].

Metabolic syndrome Much like ASD, metabolic syndrome (also known as insulin resistance syndrome and syndrome X) does not describe a single disorder but rather a cluster of interconnected health risk factors [28]. Specifically, the diagnostic criterion is the presence of at least three of the following: visceral obesity, arterial hypertension, hyperglycaemia, hypertriglyceridemia, and hypoalbuminemia [28]. MetS is recognized as a major and escalating public health challenge, chiefly in the developed world, and is thought to be caused in part by excess energy intake, and decreased energy output due to an increasingly sedentary lifestyle [36]. Metabolic syndrome is associated with an increased risk of numerous diseases and particularly notably with the development of cardiovascular disease and type 2 diabetes mellitus [2]. Approximately one third of adults in the USA suffer from MetS [25], with the prevalence of the syndrome increasing with age [25].

4.1.3. Data collection

To the best of our knowledge there are no publicly available corpora of ASD or MetS related medical literature. Hence we collected them ourselves. These are

now publicly available and can be downloaded from: <https://oa7.host.cs.st-andrews.ac.uk/>.

The data sets, their collection, and preparation are described next.

4.1.4. Raw data collection

We used the PubMed interface to access the US National Library of Medicine and retrieve abstracts and references of life science and biomedical scholarly articles. We assumed that a paper is related to ASD or MetS respectively if the terms “autism” or “metabolic syndrome” are present in its title or abstract, and collected only papers written in English. The earliest publications fitting our criteria are by Kanner [34] on ASD and by Berardinelli *et al.* [7] on MetS. We collected all matching publications up to the final one indexed by PubMed on 10th May 2015, yielding a corpus of 22,508 on ASD and 31,706 publications on MetS. We used the abstract text to evaluate our method.

4.1.5. Data pre-processing

Data collected in the manner described in the previous section comprises abstracts as freeform text. To prepare it for the type of analysis described in Section 3 we perform a series of ‘pre-processing’ steps. The goal is to remove words which are largely uninformative in any context, reduce dispersal of semantically equivalent terms, and thereafter select terms which are included in the vocabulary over which topics are learnt.

We firstly applied soft lemmatization using the WordNet[®] lexicon [42] to normalize for word inflections. No stemming was performed to avoid semantic distortion often effected by heuristic rules used by stemming algorithms. After lemmatization and the removal of so-called stop-words, we obtained approximately 2.2 and 3.8 million terms in the entire corpus when repetitions are

counted, and 37,626 and 46,114 unique terms, for the ASD and MetS corpora respectively. Constructing the vocabulary for our method by selecting the most frequent terms which explain 90% of the energy in a specific corpus resulted in ASD and MetS vocabularies containing 3,417 and 2,839 terms respectively.

4.2. Inter-topic similarity measures

Recall that in this work topics are probability distributions over a fixed vocabulary of terms. Thus the inter-topic similarity measure used to construct our temporal graph are thus similarity measures between probability distributions. Considering that the vocabulary of terms is fixed we represent each topic, say p , using a fixed length vector:

$$p = [p_1, p_2, \dots, p_{n_v}]^T \quad (13)$$

where n_v is the number of terms in the vocabulary.

For the experiments in this paper we adopted three well known measures for quantifying the similarity (or equivalently, dissimilarity) between probability distributions representing extracted topics. The first of these is the well known Hellinger distance [30]. For two discrete probability distributions, e.g. p and q representing two topics, it is defined as follows:

$$H(p, q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n_v} (\sqrt{p_i} - \sqrt{q_i})^2}. \quad (14)$$

It can be readily seen that $H()$ is symmetric and that it takes on a value between 0 and 1, with 0 signifying the greatest degree of similarity between p and q (in this case $p = q$) and 1 the least (in this case $p_i > 0 \implies q_i = 0 \wedge q_i > 0 \implies p_i = 0$).

The second similarity measure we evaluate is the Bhattacharyya coefficient

defined as [12]:

$$B(p, q) = \sum_{i=1}^{n_v} \sqrt{p_i q_i}. \quad (15)$$

As the Hellinger distance, the Bhattacharyya coefficient is symmetric and takes on a value between 0 and 1. However note that in this case it is the maximum value of 1 which is attained when there is the greatest degree of similarity between p and q (i.e. $p = q$) and 0 the least (as before in this case $p_i > 0 \implies q_i = 0 \wedge q_i > 0 \implies p_i = 0$). It is straightforward to demonstrate that the Bhattacharyya coefficient is related to the Hellinger distance as follows:

$$H(p, q) = \sqrt{1 - B(p, q)}. \quad (16)$$

Lastly, due to its widespread use we also compare the performance of the proposed algorithm using the similarity measure often erroneously referred to as Tanimoto similarity [39] (or the Jaccard similarity [24]), defined as:

$$T(p, q) = \frac{\sum_{i=1}^{n_v} p_i q_i}{\sum_{i=1}^n p_i^2 + \sum_{i=1}^{n_v} q_i^2 - \sum_{i=1}^{n_v} p_i q_i}. \quad (17)$$

As the Bhattacharyya coefficient, this measure is symmetric and takes on a value between 0 and 1, with the maximum value of 1 being attained when p and q are equal, and 0 when there is no overlap between them (i.e. $p_i > 0 \implies q_i = 0 \wedge q_i > 0 \implies p_i = 0$). In an effort to avoid perpetuating the aforementioned misnomer on the one hand while retaining some nomenclatural continuity with the existing literature, we will refer to this measure as quasi-Jaccard similarity.

4.3. Experiments and results

In this section we conduct experiments on the two corpora of scholarly literature described in Section 4.1.3, and report and discuss our results both using quantitative findings and representative qualitative examples.

4.3.1. Quantitative comparison

We started our evaluation with an experiment examining quantitative differences effected by changing different flexible parameters of the proposed method. In particular, our aim was to see how the evolving topic structure extracted by our algorithm is affected by different choices of inter-topic similarity measures (introduced in described in Section 3.2.1 and used to construct the temporal similarity graph as described in Section 3.2.2), different pruning thresholds used to infer complex structural changes, lengths of epochs used to discretize the timespan of a data corpus, and the overlap between successive epochs. As explained in the previous section, we compared three inter-topic similarity measures based on the Hellinger distance, the Bhattacharyya coefficient, and what we term the quasi-Jaccard similarity. Different combinations of epoch lengths and successive epoch overlaps are summarized in Table 1 and Table 2 for respectively the ASD and the metabolic syndrome data sets. Notice that due to the different timespans of the two data sets, different epoch lengths were used in the corresponding experiments. Similarly, epoch overlaps were adjusted (to the closest year) so as to be comparable on a relative basis so that, for example, the maximum overlap examined was approximately half of the duration of an epoch (10 : 5 vs. 6 : 3). At the other extreme we also performed experiments using no epoch overlap, as well as an additional setting per experiment with an overlap of approximately 25% of epoch length (10 : 2 and 5 : 1 in Table 1, and 6 : 1 in Table 2) .

Similarity measures To begin with, consider the results summarized in Figure 6. The figure comprises three sets of plots with each set corresponding to one of the three compared inter-topic similarity measures, and showing the number of topic births, deaths, merges, and splits per epoch, normalized by the total

Table 1. Different combinations of values of the free parameters of the proposed algorithm used for the experiments on our autism spectrum disorder abstracts data set.

Setting #	1	2	3	4	5	6
Epoch length (years)	10	10	10	5	5	5
Epoch overlap (years)	5	2	0	2	1	0

Table 2. Different combinations of values for the free parameters of the proposed algorithm used in our comparative evaluation experiments on the metabolic syndrome abstracts data set.

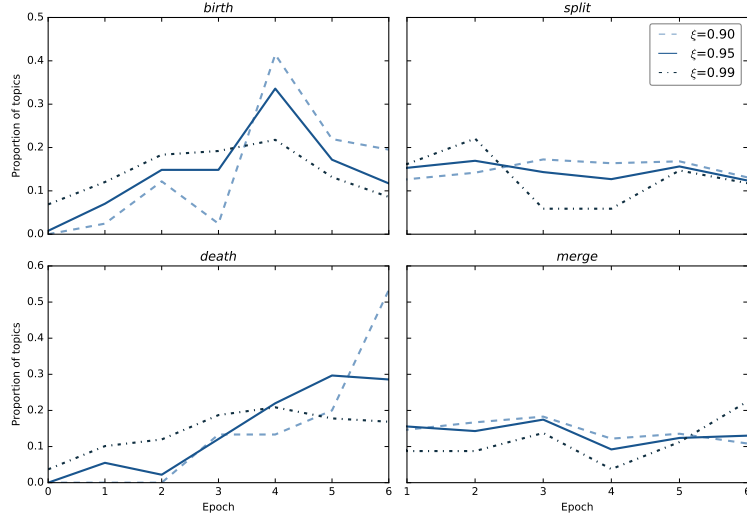
Setting #	1	2	3	4	5
Epoch length (years)	6	6	6	3	3
Epoch overlap (years)	3	1	0	1	0

number of topics inferred in that epoch. Experiments were performed using three different operating cutoff points on the empirical cumulative density function of inter-topic similarities across the initial temporal graph – the results are shown using lines of different styles, as per the plot legends. The epoch length and the amount of overlap of successive epochs were kept constant in all experiments.

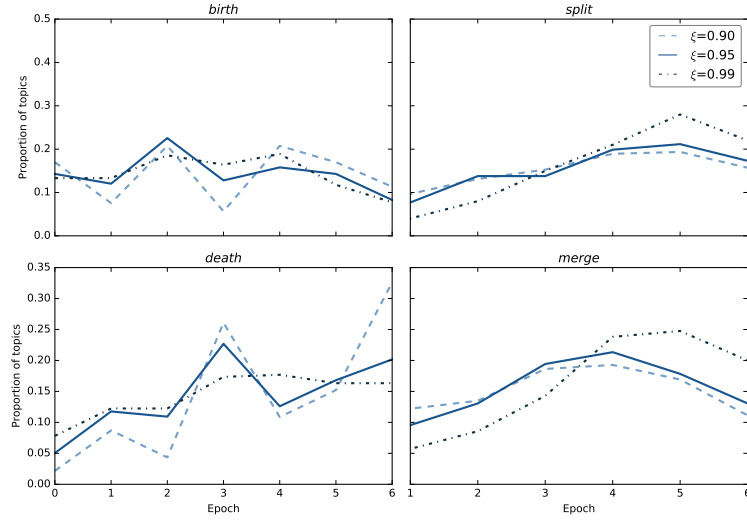
Comparing the corresponding plots across different similarity measures, it can be readily observed that the results obtained using the Hellinger distance and the Bhattacharyya coefficient are very similar. This is unsurprising considering the close relatedness of the two measures, as highlighted previously in Section 4.2 and as expressed by the expression in (16). However, at first sight the results obtained using the quasi-Jaccard index appear rather different. This raises the question which similarity measure is better (and on what grounds can such a claim be made given the fundamental lack of objective ground truth), or in a

weaker form, which similarity measure should be preferred and when i.e. if this preference should be universal, application driven, or in some sense intimated by data itself. However a closer inspection of the plots reveals interesting and important insight that obviates this daunting question. In particular it can be noticed that while there indeed are differences in behaviour between the plots corresponding to the quasi-Jaccard index and those which correspond to the Hellinger distance and the Bhattacharyya coefficient in early epochs, eventually the three converge to approximately the same terminal behaviour. The initial differences are readily explained by the well known small sample size effect – in the early stages of processing a longitudinal corpus inference relies on much fewer documents than later on, especially in the particular case considered here given that the amount of published research in ASD has been growing rapidly year after year as illustrated in Figure 7. This hypothesis is further supported by the observation that the choice of different operating points on the CDF of inter-topic similarities across the initial temporal graph, used for its subsequent pruning, also has little effect in the long run – following transient changes, the rates of different types of topic changes converge towards the same pattern.

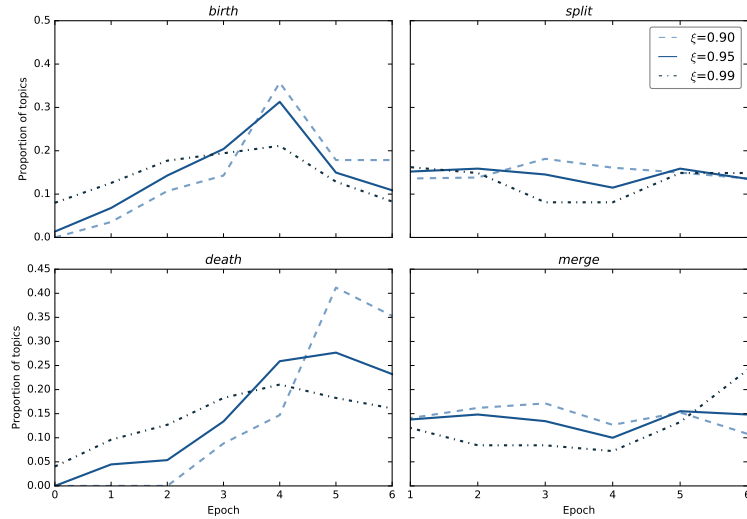
Epoch length We next examined the effects that the choice of the epoch length has on the output of our algorithm. A representative set of results is shown in Figures 8 and 9. Each figure comprises two sets of plots with each set corresponding to a particular set of algorithm parameters. The key parameter which was varied was the epoch length. For example, in Figure 8 the results obtained using the epoch length of 5 years was compared with those obtained using the epoch length of 10 years on the ASD abstracts data set. To ensure a fair comparison, successive epoch overlap was not kept constant on an absolute but rather rela-



(a) Hellinger, 10 year epochs, 5 year overlap, ASD abstracts



(b) Quasi-Jaccard, 10 year epochs, 5 year overlap, ASD abstracts



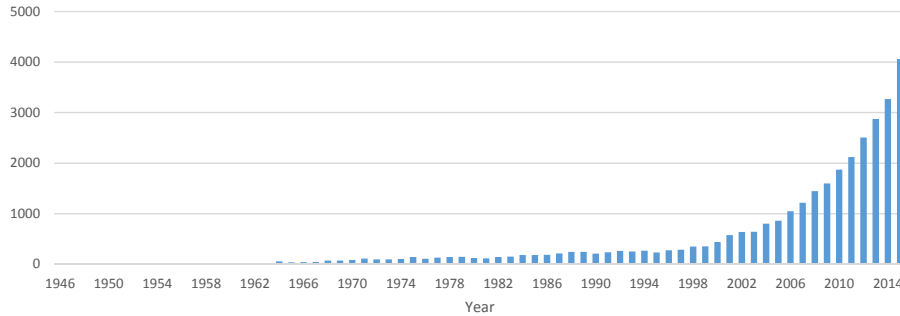
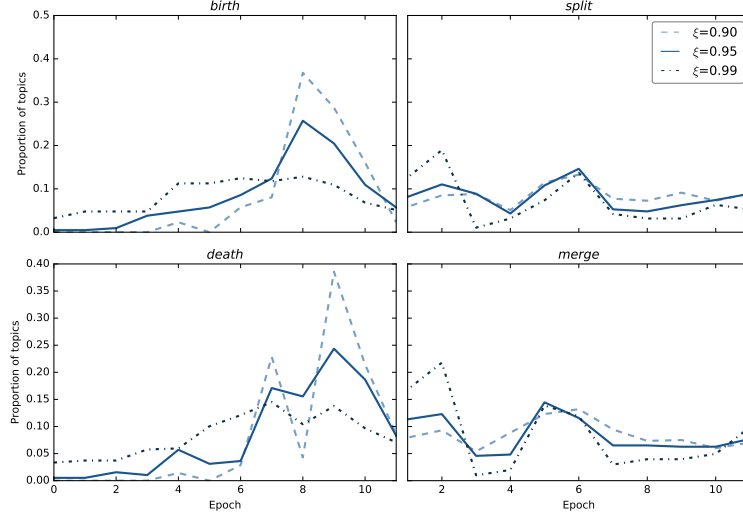


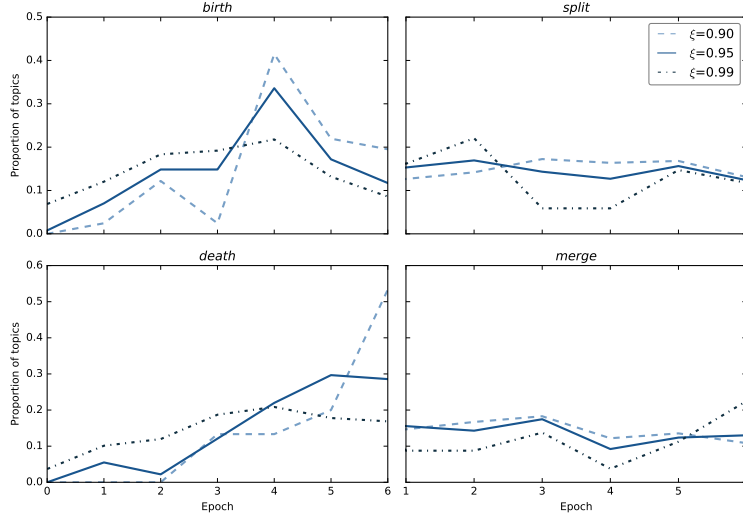
Fig. 7. The rapid rise in the rate of publications concerned with ASD. Shown is the number of publications per annum indexed by PubMed and matching our criterion for the inclusion in our ASD abstracts data set (please see Section 4.1.4).

tive basis. Specifically, the overlap is in all cases approximately 50% of the epoch length (5 : 2 and 10 : 5 in Figure 8, and 6 : 3 and 3 : 1 in Figure 9). As before each set of plots displays the number of topic births, deaths, merges, and splits per epoch, normalized by the total number of topics inferred in that epoch, with the three different operating cutoff points on the empirical cumulative density function of inter-topic similarities across the initial temporal graph shown using lines of different styles, as per the plot legends.

Successive epoch overlap Lastly we analysed the effect that the choice of successive epoch overlap has on the output of our algorithm. As in the preceding experiments we visualize a summary of the results using sets of plots of normalized rates of topic birth, death, splitting, and merging, each set corresponding to a different overlap – please see Figure 10. Even a cursory examination of the plots readily reveals that the parameter in question has a profound effect. While a consistent pattern of changes can be observed in each set of plots, the most noticeable effect is on the rate of topic death. In particular it can be seen that



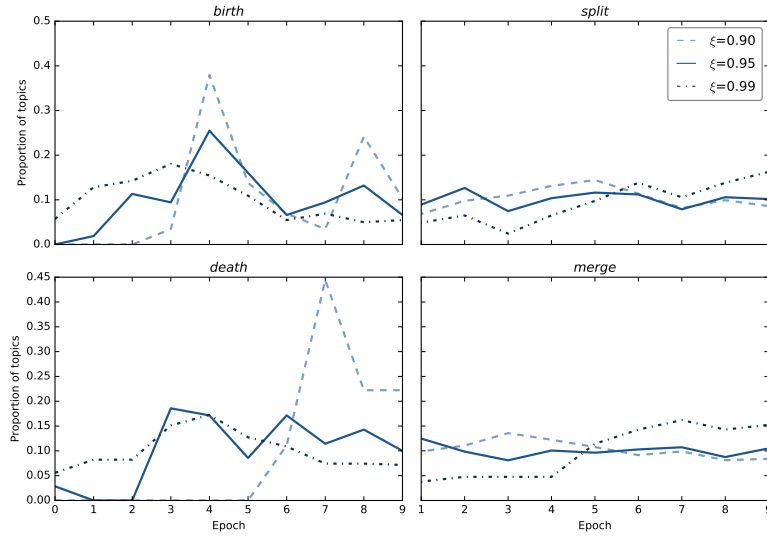
(a) Hellinger, 5 year epochs, 2 year overlap, ASD abstracts



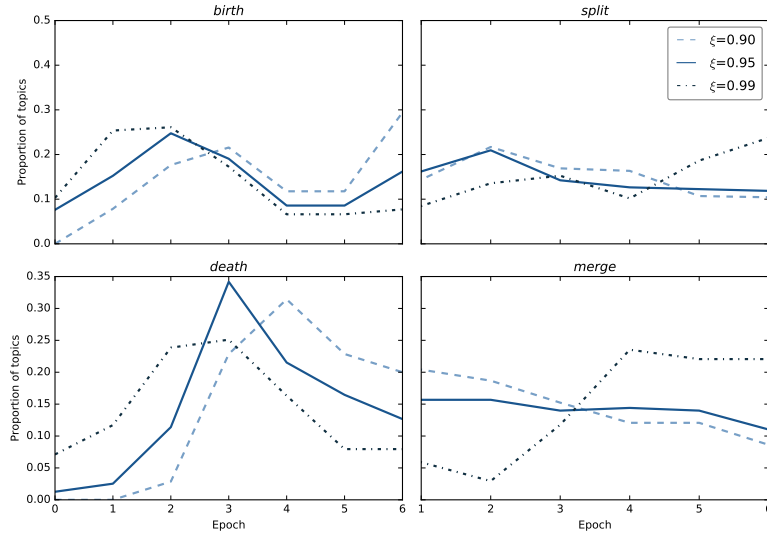
(b) Hellinger, 10 year epochs, 5 year overlap, ASD abstracts

Fig. 8. The effect of the epoch length; ASD abstracts data set.

as epoch overlap is reduced, the rate at which topics die off is increased. For example on our ASD data set using 10 year epochs, for the overlap of 5 years and the similarity CDF operating point $\zeta = 0.95$ the steady state topic death



(a) Hellinger, 3 year epochs, 1 year overlap, MetS abstracts

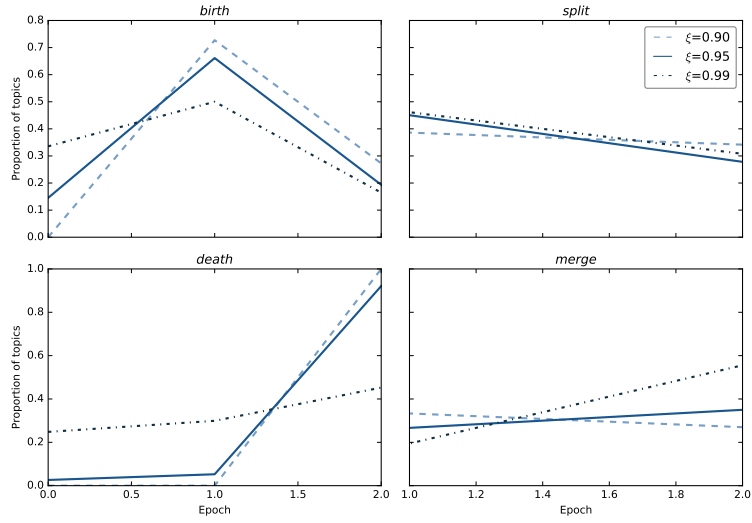


(b) Hellinger, 6 year epochs, 3 year overlap, MetS abstracts

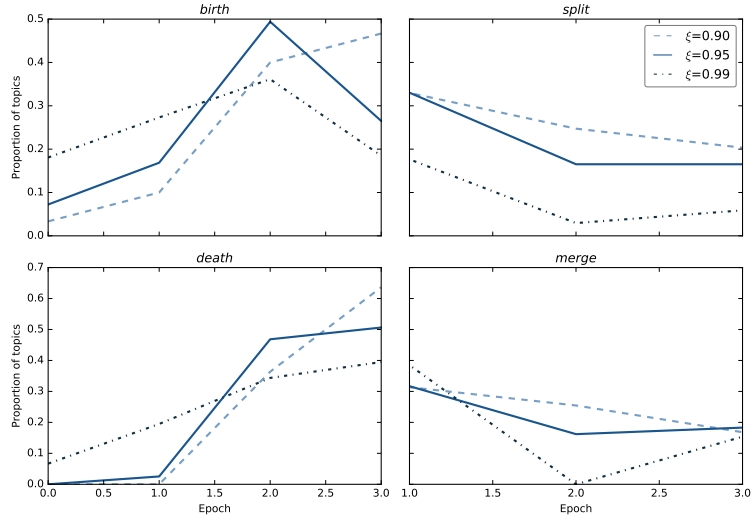
Fig. 9. The effect of the epoch length; MetS abstracts data set.

rate is approximately 0.3. It increases to approximately 0.5 as the overlap is reduced to 2 years, and then to over 0.9 for no overlap at all. In other words, in the latter case over 90% of topics exhibit no continuity of any sort (evolution, merging, or splitting), instead disappearing already within the epoch of their birth. Qualitatively this comes as no surprise – the less overlap there is between successive epochs, the less relatedness can be expected between the sets of topics extracted in successive epochs. However, qualitatively, the magnitude of the effect is rather astonishing. Considering that most of the methods described in the literature which discretize time by epochs adopt the no-overlap design, our finding provides strong and valuable evidence that the performance of these methods could be improved with little effort, merely by a slight alteration in the manner discretization is performed.

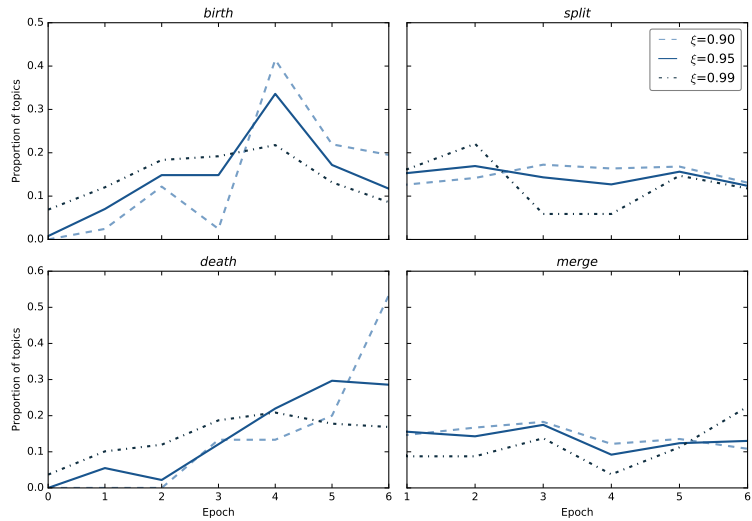
Parameter combinations and topic life expectancy In the experiments so far we examined how the topic structure of a longitudinal document corpus and the evolution of this structure over time is affected by different free parameters of the proposed algorithm. Our results suggest that our algorithm is not very sensitive to the exact choice for the value of its parameters. This behaviour is highly desirable because it obviates the need for substantial amounts of data needed to learn sensible parameter values for a particular application. The one parameter which we found to be of particular importance is the amount of overlap between successive epochs used to discretize time. In particular we found that while the precise amount of overlap is not of particular importance, the introduction of *some* overlap (25–50% of the epoch length) significantly improves the ability of our algorithm to capture temporal structural changes in the topic content of a corpus. This was most significantly demonstrated in the rate of topic death. As



(a) Hellinger, 10 year epochs, no overlap, ASD abstracts



(b) Hellinger, 10 year epochs, 2 year overlap, ASD abstracts



we noted earlier, without any epoch overlap most topics die off within the same epoch in which they are created. Hence we sought to investigate how the topic life expectancy is affected by different combinations of our algorithm’s parameter values.

In the experiments described so far we examined the normalized topic birth and death rates independently. In other words we looked at the proportion of topics which are respectively newly created (born) and which disappear (die) as a proportion of the total number of topics extracted in the corresponding epoch. This information does not provide any insight into *which* topics die off, that is, how long a specific topic has been in existence before it disappears. Here we adopt a more nuanced approach which comprises the following steps:

– **Identification of topic creation:**

We identify the epoch in which a particular topic was created as the epoch of its birth *or* the epoch in which the topic is created through the splitting or the merging of topics from the previous epoch.

– **Topic tracking:**

Following the creation of a topic we track it in the context of complex changes of its topic environment by considering its natural descendent to be its child in our temporal topic graph, with the highest degree of inter-topic similarity across all siblings. A topic is considered extant as long as it has any descendants.

– **Identification of topic death:**

Finally, we identify the epoch of a topic’s death as the epoch in which the topic no longer has any descendants (children in our temporal topic graph).

Adopting this methodology we analyse the distribution of the life expectancy

of a topic across our corpus and the effects of different algorithm parameters. Using the operating point corresponding to $\zeta = 0.95$ on the inter-topic similarity CDF, we performed experiments using the six settings summarized in Table 1. Our results are shown in Table 3.

Lastly, we examined the manner in which topics extracted by our algorithm cease to exist. In particular, for each new topic (where “new” in this context is taken to mean that a topic is either newly born, as defined previously, or that it is created by splitting or merging of topics from the previous epoch) we compute the time until it either dies (i.e. has no offspring in the following epoch), or splits or merges *without any of its children* having it as the sole parent. The condition that none of a topic’s children have the original topic as the sole parent is important because if there are such children, they can be seen to be evolved successors. If there are multiple children which have the ancestral topic as the sole parent then we track the lifetime of the topic down that path which leads to the longest topic lifetime. Our results are summarized in Table 3 and the plots in Figure 11. In Table 3 it is important to observe the lack of sensitivity of our method to its specific settings. From the plots in Figure 11 it is interesting to notice that most of the topics cease to exist already in the first epoch. This observation provides potentially useful insight into the choice of the temporal analysis scale. As expected, increasing the pruning threshold ζ , which has the effect of decreasing inter-epoch topic linkage, results in an increased proportion of outright topic deaths. Lastly, it is insightful to observe that following the first epoch, for a fixed specific pruning threshold the proportion of topic deaths decreases.

Table 3. Average lifespan of topics (in epochs) extracted from our autism spectrum disorder abstracts data set. *For full detail of different settings see Table 1.

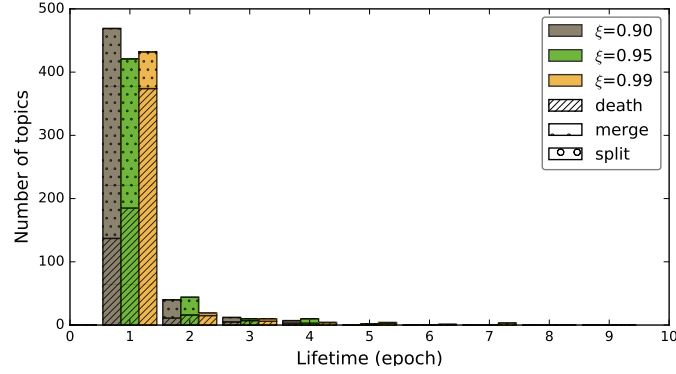
Epoch length	10 years				5 years	
Setting number*	1	2	3	4	5	6
Hellinger	1.28	1.27	1.21	1.38	1.30	1.27
Bhattacharyya	1.31	1.30	1.22	1.39	1.31	1.26
Pseudo-Jaccard	1.30	1.23	1.15	1.23	1.28	1.17

4.3.2. Qualitative results

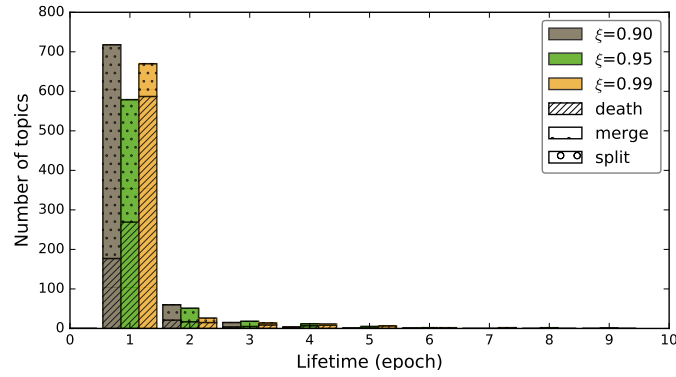
In the previous section we described an extensive set of experiments providing insight into the role that different free parameters of our algorithm have on its output. Importantly we demonstrated that within a wide range of what may be described as reasonable choices for the parameter values, the proposed method exhibits a high degree of robustness. Following these encouraging quantitative findings, using our (human) higher level semantic understanding of the corpora used, we now examine if the output of our algorithm is meaningful and ultimately useful.

Case study 1: ASD and genetics. While the exact aetiology of the ASD is still poorly understood, the existence of a significant genetic component is beyond doubt [41]. Work on understanding complex genetic factors affecting the development of autism, which possibly involve multiple genes which interact with each other and the environment, is a major theme of research and as such a good case study on which the usefulness of the proposed method can be illustrated.

We started by identifying the topic of interest as that with the highest probability of the terms “gene” or “genetic” conditioned on the topic, and tracing



(a)



(b)

Fig. 11. A statistical summary of the manner in which topics cease to exist. The plots correspond to our ASD data set, analysed using the Bhattacharyya similarity measure, and (a) 10 year epochs with a 5 year overlap, and (b) 5 year epochs with a 2 year overlap.

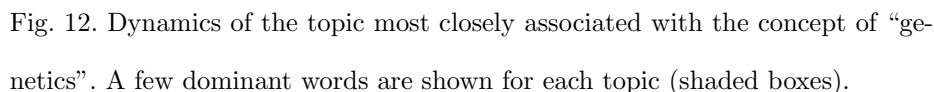
it back in time to the epoch in which it originated. This led to the discovery of the relevant topic in the epoch spanning the period 1986–1991. Figure 12 shows the evolution of this topic from 1992 revealed by our method (due to space constraints only the most significant parts of the similarity graph are shown; minor changes to the topic before 1992 are also omitted for clarity, as indicated by the

dotted line in the figure). Each topic is labelled with its first few dominant terms. The following interpretation of our findings is readily apparent. Firstly, in the period 1992–1997, the topic is rather general in nature. Over time it evolves and splits into topics which concern more specific concepts (recall that such splitting of topics cannot be captured by any of the existing methods). For example by the epoch 2002–2007 the single original topic has evolved and split into four topics which concern the following:

- The relationship between mutations in the gene `mecp2` (essential for normal functioning of neurone), and mental disorders and epilepsy (it is estimated that one third of ASD individuals also have epilepsy),
- Gene alternations, for example the duplication of `15q11--13` and deletion of `16p11.2` both of which are associated with ASD,
- Genetic linkage association analysis and heritability of autism, and
- Observational work on autistic twins and probands with siblings on the spectrum.

Our framework also allows us to look ‘back’ in time. For example, by examining the topics that the 1992 genetics topic originate from we discovered that the topic evolved from the early concept of “infantile ASD” [34].

Case study 2: ASD and vaccination. For our second case study we chose to examine research on the relationship between ASD development and vaccination. This subject has attracted much attention both in the research community, as well as in the media and the general public. The controversy was created with the



As in the previous example, we begin by identifying the topic with the highest probability of the terms “vaccine” and “vaccination” conditioned on the topic, and tracing it back to the epoch in which it first emerged. Again, a single topic was readily identified, in the epoch spanning the period 1996–2001. Notice that this is consistent with the publication date of the first relevant publication by Wakefield [57]. The evolution of the topic is illustrated in Figure 13 in the same way as in the previous section. It can be seen that the original topic concerned the subjects initially brought to attention such as “measles”, “vaccine”, and “autism”. In the subsequent epoch, when the original claim was still thought

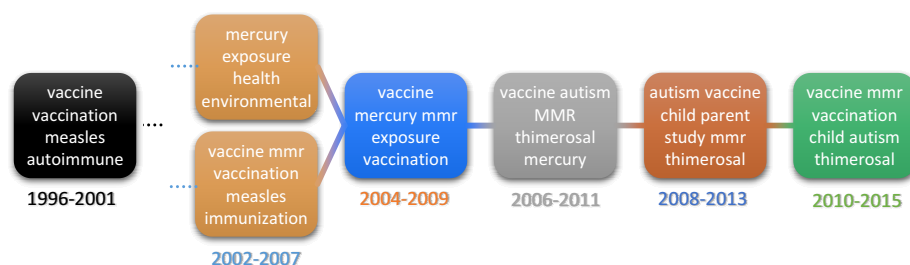


Fig. 13. Dynamics of the topic most closely associated with the concept of “vaccination”. Notwithstanding the rejection of any link between vaccination and autism, this topic remains active albeit in a form which evolved over time.

to have credibility, the topic evolves and splits into numerous others mirroring research directions taken by various researchers. Following this period and the revelations of its fraudulence, the topic assumes mainly single-threaded evolution, at times incorporating various originally separate ideas. For example observe the independent emergence of the term “mercury”. Though initially unrelated to it this topic merges with the topic that concerns vaccination which can be explained by the widely publicized thiomersal (vaccine preservative) controversy (again note that such merging of topics cannot be captured by the existing methods). Although rejected by the medical community due to a lack of evidence, this topic can be seen as persisting to date.

Case study 3: MetS and plasma fatty acids. As noted earlier MetS is highly associated with the risk of developing type 2 diabetes and cardiovascular diseases, and is characterized by insulin resistance, abdominal obesity, and high blood pressure, all of which are intimately linked with dyslipidemia and elevated plasma fatty acid levels. In this case study we sought to investigate patterns associated with topics concerning this aspect of MetS.

As in the previous two case studies we began by identifying the topics with the highest probability of the relevant terms (in this case “acid” and “fatty”) conditioned on the topic, and tracing it back to the epoch in which it first emerged. For the sake of clarity of visualization we focus on the period starting with the 1999–2002 epoch which shows some of the most interesting dynamics. The evolution of topic of interest in this period is shown in Figure 14. Examples of illustrative observations based on this section of our similarity graph include the following:

- The 2001–2004 epoch gives birth to a topic dominated by the terms “acid” and “uric”. This topic merges a topic concerned with dietary fat, thereby resulting in a topic associated with hyperuricemia in the following epoch. Both the semantics of the extracted topics themselves as well as the aforementioned dynamics are readily interpreted from and supported by the literature on MetS – the relationship between dietary fat (and types thereof) intake and and plasma uric acid levels has been a topic of significant research interest [32].
- The broad topic concerning fatty acid metabolism present in the 2005–2008 epoch can be seen to branch into three distinct research directions on the accumulation of free fatty acids, controlled clinical trials in murine models, and hyperuricemia and gout. This pattern is again congruent with the actual development of research in the field in recent years: the role of free fatty acid accumulation has been attracting an increasing amount of research attention [49], murine models have been widely used to study specific aspects of MetS in controlled conditions unfeasible with human subjects [40], and there is an accumulating body of evidence supporting a causal link between MetS

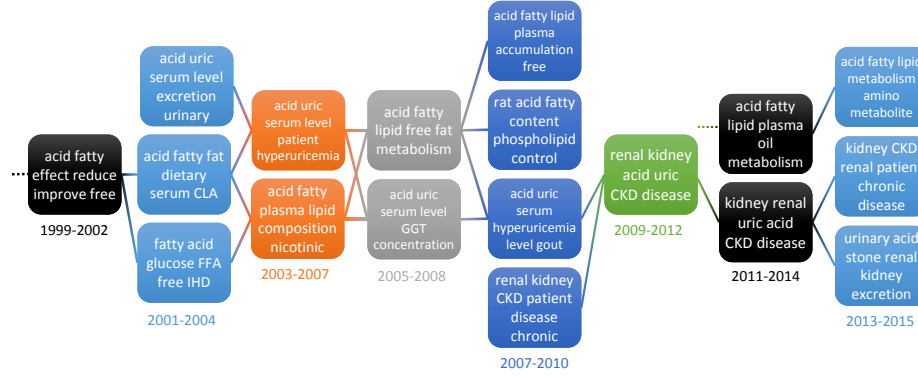


Fig. 14. Dynamics of the topic most closely associated with the concept of fatty acids (i.e. in the context of our vocabulary the terms “acid” and “fatty”). Fatty acid metabolism plays a key role in the metabolic syndrome.

and the increasing incidence of gout (historically known as “the rich man’s disease”) in the Western world [17].

5. Summary and conclusions

This paper focused on the problem of modelling and extracting the topic structure of a longitudinal document corpus over time. The approach we described starts with a discretization of time into epochs which may overlap. Then, using the approximation that the topic structure within each epoch is temporally locally static, the aforesaid structure is modelled and extracted using a hierarchical Dirichlet process. Finally, the evolution of the topic structure over time is captured using a temporal graph underlain by an inter-topic similarity measure. The graph, initially populated by edges between all pairs of topics in two consecutive epochs is pruned automatically and the result used to infer complexity structural changes over time which the existing methods in the literature cannot

model, including the emergence and disappearance of topics and their evolution over time, as well as the merging and splitting of an arbitrary number of topics.

The proposed framework was evaluated extensively on two large real-world data sets of abstracts of scientific papers, one concerning the autism spectrum disorder and the other the metabolic syndrome. This data was collected by ourselves, and made free for public use. Our detailed quantitative analysis of the effects that the free parameters of the proposed method have on its performance revealed a number of important insights. We found that within a wide range of parameter values our algorithm was little affected by the specific value choices. Another important finding, the significance of which extends further than the scope of the proposed algorithm, is that in the discretization of time into epochs it is important that successive epochs overlap. The significantly inferior performance observed with non-overlapping epochs has immediate consequences for the interpretation of previous work and the findings reported in the literature, suggesting a simple and immediate way of enhancing the performance of any algorithm which did not adopt the use of overlapping epochs. Lastly, on several case studies highly relevant to the currently popular directions of research on ASD and MetS, our algorithm’s output was analysed qualitatively, and shown to capture well the actual developments in these fields.

References

- [1] Autism spectrum disorder fact sheet. *American Psychiatric Publishing*, 2013.
- [2] K. G. M. M. Alberti, R. H. Eckel, S. M. Grundy, P. Z. Zimmet, J. I. Cleeman, K. A. Donato, J.-C. Fruchart, W. P. T. James, C. M. Loria, and S. C. Jr Smith. Harmonizing the metabolic syndrome. *Circulation*, 120(16):1640–1645, 2009.
- [3] O. Arandjelović and R. Cipolla. Face set classification using maximally probable mutual

- modes. *In Proc. IAPR International Conference on Pattern Recognition*, pages 511–514, 2006.
- [4] Corey W Arnold, Suzie M El-Saden, Alex AT Bui, and Ricky Taira. Clinical case-based retrieval using latent topic analysis. *AMIA*, 2010:26, 2010.
- [5] Corey W Arnold and William Speier. A topic model of clinical reports. *SIGIR*, pages 1031–1032, 2012.
- [6] A. J. Baxter, T. S. Brugha, H. E. Erskine, R. W. Scheurer, T. Vos, and J. G. Scott. The epidemiology and global burden of autism spectrum disorders. *Psychological Medicine*, 45(3):601–613, 2015.
- [7] W Berardinelli, J Gabriel Cordeiro, Danilo de Albuquerque, and Antonio Couceiro. A new endocrine-metabolic syndrome probably due to a global hyperfunction of the somatotrophin. *Acta endocrinologica*, 12(1):69–80, 1953.
- [8] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. *In Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1:550–562, 2015.
- [9] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 1354–1361, 2015.
- [10] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Data-mining Twitter and the autism spectrum disorder: a pilot study. *In Proc. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, pages 349–356, 2014.
- [11] A. Beykikhoshk, O. Arandjelović, D. Phung, S. Venkatesh, and T. Caelli. Using Twitter to learn about the autism community. *Social Network Analysis and Mining*, 5(1):5–22, 2015.
- [12] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–109, 1943.
- [13] David M Blei, K Franks, Michael I Jordan, and I Saira Mian. Statistical modeling of

- biomedical corpora: mining the *Caenorhabditis* genetic center bibliography for genes related to life span. *BMC Bioinformatics*, 7(1):250, 2006.
- [14]David M Blei and John D Lafferty. Dynamic topic models. *ICML*, pages 113–120, 2006.
- [15]David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.
- [16]J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: how humans interpret topic models. *Advances in Neural Information Processing Systems*, pages 288–296, 2009.
- [17]H. K. Choi, E. S. Ford, C. Li, and G. Curhan. Prevalence of the metabolic syndrome in patients with gout: the third national health and nutrition examination survey. *Arthritis Care Res*, 57(1):109–115, 2007.
- [18]J. T. Danial and J. J. Wood. Cognitive behavioral therapy for children with autism: Review and considerations for future research. *Journal of Developmental & Behavioral Pediatrics*, 34(9):702–715, 2013.
- [19]Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
- [20]Avinava Dubey, Ahmed Hefny, Sinead Williamson, and Eric P Xing. A nonparametric mixture model for topic modeling over time. *SDM*, pages 530–538, 2013.
- [21]F. J. Dyson. Is science mostly driven by ideas or by tools? *Science*, 338(6113):1426–1427, 2012.
- [22]A. Einstein and L. Infeld. *The evolution of physics: the growth of ideas from early concepts to relativity and quanta*. Cambridge University Press, 1961.
- [23]Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.
- [24]M. A. Fligner, J. S. Verducci, and P. E. Blower. A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, 44(2):110–119, 2002.
- [25]E. S. Ford, W. H. Giles, and W. H. Dietz. Prevalence of metabolic syndrome among us adults: findings from the third National Health and Nutrition Examination Survey. *JAMA*, 287(3):356–359, 2002.

- [26]D. E. Gray. Perceptions of stigma: The parents of autistic children. *Sociology of Health & Illness*, 15(1):102–120, 1993.
- [27]Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.
- [28]S. M. Grundy, H. B. Jr Brewer, J. I. Cleeman, S. C. Jr Smith, and C. Lenfant. National Heart, Lung, and Blood Institute, American Heart Association: Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition. *Circulation*, 109:433–438, 2004.
- [29]J. W. Harrington, L. Rosen, A. Garnecho, and P. A. Patrick. Parental perceptions and use of complementary and alternative medicine practices for children with autistic spectrum disorders in private practice. *Journal of Developmental & Behavioral Pediatrics*, 27(2):S156–S161, 2006.
- [30]E. Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.
- [31]Thomas Hofmann. Probabilistic latent semantic indexing. *SIGIR*, pages 50–57, 1999.
- [32]L. C. Hudgins, M. Hellerstein, C. Seidman, R. Neese, J. Diakun, and J. Hirsch. Human fatty acid synthesis is stimulated by a eucaloric low fat, high carbohydrate diet. *J Clin Invest*, 97(9):2081, 1996.
- [33]A. Hviid, M. Stellfeld, J. Wohlfahrt, and M. Melbye. Association between thimerosal-containing vaccine and autism. *The Journal of the American Medical Association*, 290(13):1763–1766, 2003.
- [34]Leo Kanner. Irrelevant and metaphorical language in early infantile autism. *American Journal of Psychiatry*, 103(2):242–246, 1946.
- [35]Vinod D Kumar and Hannah Jane Tipney. *Biomedical Literature Mining*. Springer, 2014.
- [36]T. A. Lakka, D. E. Laaksonen, H. M. Lakka, N. Männikkö, L. K. Niskanen, R. Rauramaa, and J. T. Salonen. Sedentary lifestyle, poor cardiorespiratory fitness, and the metabolic syndrome. *Med Sci Sports Exerc*, 35(8):1279–1286, 2003.
- [37]S. E. Levy, D. S. Mandell, and R. T. Schultz. Autism. *Lancet*, 374(9701):1627–1638, 2009.
- [38]J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

- [39]A. H. Lipkus. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*, 26(1):263–265, 1999.
- [40]B. Mackness, R. Quarck, W. Verreth, M. Mackness, and P. Holvoet. Human paraoxonase-1 overexpression inhibits atherosclerosis in a mouse model of metabolic syndrome. *Arterioscler Thromb Vasc Biol*, 26(7):1545–1550, 2006.
- [41]J H Miles. Autism spectrum disorders – a genetics review. *Nature*, 13(4):278–294, 2011.
- [42]G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [43]Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.
- [44]Lu Ren, David B Dunson, and Lawrence Carin. The dynamic hierarchical Dirichlet process. *ICML*, pages 824–831, 2008.
- [45]Daniel R Rhodes, Jianjun Yu, Kalyan Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pander, and Arul M Chinnaiyan. A cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, 2004.
- [46]Sylvia Richardson and Peter J Green. On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.
- [47]W. Rieutort-Louis and O. Arandjelović. Bo(V)W models for object recognition from video. *In Proc. International Conference on Systems, Signals and Image Processing*, pages 89–92, 2015.
- [48]F B Rogers. Medical subject headings. *Bull Med Libr Assoc*, 51:114–116, 1963.
- [49]A. Seppälä-Lindroos, S. Vehkavaara, A. M. Häkkinen, T. Goto, J. Westerbacka, A. Sovijärvi, J. Halavaara, and H. Yki-Järvinen. Fat accumulation in the liver is associated with defects in insulin suppression of glucose production and serum free fatty acids independent of obesity in normal men. *J Clin Endocrinol Metab*, 87(7):3023–3028, 2002.
- [50]Jayaram Sethuraman. A constructive definition of Dirichlet priors. Technical report, DTIC Document, 1991.
- [51]Burr Settles. ABNER: an open Source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

- [52] Matthew S Simpson and Dina Demner-Fushman. *Mining text data*, chapter Biomedical text mining: a survey of recent progress, pages 465–517. 2012.
- [53] D R Swanson. Undiscovered public knowledge. *Library Quarterly*, 56(2):103–118, 1986.
- [54] Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *Advances in neural information processing systems*, pages 1353–1360, 2006.
- [55] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.
- [56] D. Trembath, S. Balandin, and C. Rossi. Crosscultural practice and autism. *Journal of Intellectual and Developmental Disability*, 4(30):240–242, 2005.
- [57] A J Wakefield, S H Murch, and A Anthony. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*, pages 637–641, 1998. (Retracted).
- [58] Chong Wang, David Blei, and David Heckerman. Continuous time dynamic topic models. *UAI*, pages 579–586, 2008.
- [59] Xuerui Wang and Andrew McCallum. Topics over time: a non-Markov continuous-time model of topical trends. *SIGKDD*, pages 424–433, 2006.
- [60] Z. Warren, M. L. McPheeters, N. Sathe, J. H. Foss-Feig, A. Glasser, and J. Veenstra-VanderWeele. A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics*, 127(5):e1303–e1311, 2011.
- [61] Yonghui Wu, Mei Liu, W Zheng, Zhongming Zhao, and Hua Xu. Ranking gene-drug relationships in biomedical literature using latent Dirichlet allocation. *Pacific Symposium on Biocomputing*, pages 422–433, 2012.
- [62] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical Dirichlet processes for multiple correlated time-varying corpora. *SIGKDD*, pages 1079–1088, 2010.